DOCUMENT RESUME

ED 448 227 UD 033 884

AUTHOR	Webb, Norman L.
TITLE	Embedded Research in Practice: A Study of Systemic Reform in Milwaukee Public Schools.
SPONS AGENCY	Joyce Foundation, Chicago, IL.; Helen Bader Foundation, Milwaukee, WI.
PUB DATE	2000-04-00
NOTE	21p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Accountability; Change Strategies; *Educational Change; *Educational Research; Elementary Secondary Education; Public Schools; Research Methodology; *Student Evaluation; Urban Schools
IDENTIFIERS	*Milwaukee Public Schools WI; *Reform Efforts

ABSTRACT

This paper describes the Study of Systemic Reform in Milwaukee Public Schools (MPS), an embedded research project that crossed the lines between objectivity and subjectivity, technical assistance and evaluation, and qualitative and quantitative research. The project created a collaboration between researchers at the Wisconsin Center for Education Research and staff at MPS to develop a higher level of analytic and management capacity for shaping and guiding a set of reform efforts. The embedded research was guided by a theory of systemic reform that required the alignment of system components within the district. The research also incorporated system and school accountability and assessment principles, including consequential validity and the use of multiple measures. Inputs from the district included newly adopted standards and grade level expectations, an assessment system that incorporated a variety of measures and that drove instruction (but did not have the resources to develop psychometrically sound instruments); a strong Middle School Principal's Collaborative and coherence in the middle grades; a newly adopted district-wide mathematics curriculum for the middle grades; a data warehouse system under development; and a new administration that is decentralizing the district by moving nearly all decision responsibility to the schools. (Contains 14 references.) (SM)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 448 227

EMBEDDED RESEARCH IN PRACTICE: A STUDY OF SYSTEMIC REFORM IN MILWAUKEE PUBLIC SCHOOLS

Norman L. Webb
Wisconsin Center for Education Research
School of Education
University of Wisconsin-Madison

Paper presented at the American Educational Research Association Annual Meeting held in New Orleans, Louisiana, April 24-28, 2000.

BEST COPY AVAILABLE

UD033884

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
N.L. Webb
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Embedded Research in Practice:
Center for the Study of Systemic Reform in Milwaukee Public Schools[1]

Norman L. Webb

Embedded research, as a methodology, is akin to design experiments (Clune, 2000; Brown, 1992) and action research, but is distinct from these modes of inquiry in very specific ways. The type of inquiry we have employed on the Study of Systemic Reform in Milwaukee Public Schools crossed lines that distinguish between objectivity and subjectivity, technical assistance and evaluation, and qualitative and quantitative research. We developed a new term for this kind of hybrid inquiry because 1) existing terms did not quite fit what we were doing and 2) we sought to raise people's awareness of the fact that doing research in a systemic reform context can require new roles for a researcher or evaluator (Century, 1999).

An important goal for the Study of Systemic Reform in Milwaukee Public Schools, funded by the Joyce Foundation and Helen Bader Foundation, was to form a collaboration between researchers at the Wisconsin Center for Education Research (WCER) and staff at MPS to develop a "higher level of analytic and management capacity for shaping and guiding a set of exciting and ambitious reform efforts" (Clune & Webb, 1997, p. 1). The project was designed to serve the interests both of the district in improving its capacity and of the researchers in extending their knowledge of how systemwide change can be advanced in a large urban district. We identified three purposes (Clune & Webb, 1997, p. 2):

1. Generating useful knowledge and recommendations for policy in the district;
2. Allowing impartial observers, funding agencies, and managers to understand the system and its performance at a deeper level; and,
3. Imparting analytical capacity to the district so that the role of the Center could be phased out, or reduced, after a period of years.

The first and third purposes, which involve technical assistance, are directed toward the interests of the district and the generation of knowledge within the district for improved policy making. The second purpose is directed towards the researchers' interest in understanding more about systemic reform. As distinct from doing a research study *on the district*, we are engaged in working *with the district*. These multiple purposes and perspectives coincide with design experiment, but clearly separate our work from experimental research.

Theory is essential for guiding embedded research. As researchers, we came to the study influenced heavily by the perspective on systemic reform advanced by Smith

---

[1] The research reported in this report was supported by grants from the Joyce Foundation and the Helen Bader Foundation. The opinions, findings, and conclusions that are expressed in this report do not necessarily reflect those of the supporting foundations or the Wisconsin Center for Education Research.

and O'Day (1991) and the National Science Foundation (Zucker, Shields, Adelman, Corcoran, & Goertz, 1998). Our understanding of the theory of systemic reform was that systemic policy is the most promising method of sustaining major gains in student achievement on a continuous basis over the long run. This theory, succinctly stated by Clune (1998), is represented by a continuous causal sequence: SR→ SP → SC → SA, where SR = systemic reform, SP = systemic policy, SC = systemic curriculum, and SA = student achievement that reflects the curriculum. Even though we were guided by a theory of systemic reform, we understood that systemic reform as an approach to large-scale intervention still was an unproven theory (Heck & Webb, 1997).

Ann Brown (1992) developed design experiments that drew heavily on learning theory. Her experiment was to design implementation of cognitive learning theory in a classroom setting. Through the design process, she advanced her understanding and that of others about what is needed to effect change in teaching and to establish the validity of the learning theory. How we have used theory in embedded research differs somewhat in degree from how Brown employed theory in design experiments. Whereas Brown used learning theory *to implement* and study instruction in classrooms, we are more engaged in *refining and validating* a theory for large-scale change. Both embedded research and design experiments inform theory, but the former emphasizes theory building, whereas the latter emphasizes theory refinement. As indicated by Chen (1990) in his explanation of theory-driven evaluation, design experiments emphasize *prescriptive theory* that prescribes what ought to be done and embedded research emphasizes *descriptive theory* that describes and explains what is.

In addition to the role of theory as applied to research, embedded research is distinct from design experiments simply because of the magnitude of the study. Brown successfully employed design experiments in a classroom setting where she was the primary researcher. We are applying embedded research in a large urban district. When we began our research in 1998, Milwaukee Public Schools was the nation's fifteenth largest school district. Approximately 100,000 students were enrolled in over 150 schools. The student population consisted of 50 percent African American, 25 percent Caucasian, 11 percent Hispanic, 11 percent Asian, one percent Native American, and one percent other. About 65 percent of the students received free lunch. The district employed over 9,000 people, 6,000 of whom were teachers. For us as researchers to even assume we could implement or design an implementation intervention would simply be naive. As in any large urban district, multiple interventions were in effect, including Title I, Sage Program (reduce class sizes for primary grades), one of NSF's Urban Systemic Initiatives, Data- Driven Decision Making Seminars, Project Seed, Goals 2000 Planning, Target Teach (a reading and mathematics intervention program), P-5 (preschool to grade 5 intervention for economically disadvantage elementary school students), and over 50 more (Office of Research & Assessment, 1998). As researchers, we also did not have any authority or even direct access to the superintendent or Board of School Directors, who have the major responsibility for setting policy and articulating the vision for the district.

Clune's model of embedded research (Clune, 2000) depicts change theory and the inputs that feed into building understanding of the system and the proximate path of

systemic reform. One premise of systemic reform is that the major components of an education system must work together to guide the process of helping students achieve higher levels of understanding (Smith & O'Day, 1991; Zucker et al., 1998; Webb, forthcoming). Policy makers and educators recognize that if system components are not aligned, the system will be fragmented, will send mixed messages, and will be less effective (CPRE, 1991; Newmann, 1993). For example, the systemic initiatives program of the National Science Foundation (NSF) is directed toward states, districts, and regions setting ambitious goals for student learning that are based upon a coherent policy system. The Improving America's Schools Act explicated how assessments are to relate to standards: " . . . such assessments (high quality, yearly student assessments) shall . . . be aligned with the State's challenging content and student performance standards and provide coherent information about student attainment of such standards . . ." (U.S. Congress, 1994, p. 8). Similarly, the U.S. Department of Education's explanation of the Goals 2000: Educate America Act and the Elementary and Secondary Education Act (which includes Title I) indicated alignment of curriculum, instruction, professional development, and assessments as key performance indicators for states, districts, and schools striving to meet challenging standards.

Because of the multiplicity of challenges in studying a large urban school district and systemic reform, we assembled a multidisciplinary research team. This team included persons with a background in education policy, curriculum, professional development, special education, assessment, evaluation, and econometrics. The researchers' interests served as a starting point for inquiry in the district. But as the project progressed and the district's priorities shifted, the research studies gravitated toward satisfying the immediate needs within the district while building on the perspective and expertise of the different researchers on the team.

Context

In order to better understand the evolution of our embedded research over the course of the study, it is helpful to be aware of some of the context. In February, 1996, the Board approved a plan that required the district to develop Middle School Proficiencies for all grade 8 students. In April, 1998, Dr. Alan Brown was appointed the superintendent of Milwaukee Public Schools. Under his administration, and building on work of the prior administrations, the district launched an aggressive standards-based reform. In November, 1998, the Board of School Directors approved new curriculum standards in mathematics, communications, science, and social studies. The mathematics and science standards incorporated grade-level expectations that had been developed over a period of at least five or six years. Beginning with those in 1999-2000, grade 8 students were required to demonstrate an acceptable level of accomplishment in communication, mathematics, science, and research in order to be promoted to grade 9.

In April, 1999, one year after the initiation of our study of systemic reform in MPS, the newly-elected Board of School Directors dismissed Superintendent Brown and named Dr. Spence Korte, a long-time successful principal from the district, superintendent. Superintendent Korte replaced many district staff in key administrative

positions and initiated a new strategic planning process. An issue that immediately emerged was decentralization of the district's administration by shifting a large proportion of the district's budget to the control of the schools.

Over the summer of 1999, the state legislature and governor approved a new high school graduation test to go into effect during the 2003-2004 school year, along with accountability requirements for grades 4 and 8 promotion. According to these requirements, districts in the state are to use three criteria to decide on students' promotion and graduation—state assessments, academic performance, and teacher recommendations. Because of Wisconsin's strong tradition of local control, the legislation is very permissive and allows districts significant latitude in specifying what the requirements should be for each criterion; it also gives parents the option of taking their children out of a test. One expectation of the legislation is that if a district uses the state graduation test, then the district has to adopt the state's curriculum standards.

From August to November, 1999, the district leadership was concerned with appointing people to fill positions and engaged in a strategic planning process. Over that four-month period, our project faced a moratorium on any direct activities within the district. The moratorium on our collection of new data at the beginning of the 1999-2000 school year was helpful in giving us more time to analyze data we had already collected, but it impeded any progress we could make in advancing our research and helping the district develop its analytic capacity. However, our growing knowledge of the district and the research we had performed helped immensely when our work began with a new direction in January, 2000.

At a December, 1999, meeting, we were introduced to the district's strategic planning process. The deputy superintendent, a former middle school principal, felt our project could be most helpful in working with the Middle School Collaborative, a group of the middle school principals who had been meeting regularly and working together to resolve issues related to the middle school proficiencies and curriculum.

At the beginning of the new calendar year and with the direction of the deputy superintendent, our work was focused on helping the district in three areas. One area was to assist the district in streamlining its assessment system. A second area was to study the effectiveness of the middle grades proficiencies and the influence these were having on instruction. The direction of this inquiry was situated in the context of our working more closely with the Middle School Principals Collaborative. This group had been functioning very well and represented the direction the district wanted to move toward putting decision making in the hands of the principals. A third area was to help define better accountability criteria for the first district charter school. These criteria could become the model for specifying accountability criteria for other schools. We also continued to interact with the Technical Services unit to gain access to data that could be used to monitor the district's process in improving student learning.

Whereas in the first year of the project, we had been guided towards research on the district's accountability system, alignment among standards, assessment, and

instruction, and information systems, beginning in January 2000, the project narrowed its focus to the assessment system, middle school proficiencies, and related accountability issues. Concurrently, the project continued to gain access to data from the district that could be used to conduct or demonstrate a value-added analysis of student achievement.

Examples of Embedded Research

Alignment Study

The project has taken unanticipated turns and has had to adjust to the realities of working within an urban district. One such reality is the abrupt change in leadership. Of those with whom we had worked under the Brown administration, only a very few remained in their positions in the Korte administration. This meant that whatever progress we had made in developing trust and setting plans had to be renegotiated. However, because the research we engaged in at the onset of the project had theoretical underpinnings and was designed in part to increase our knowledge of the district, we were able, with the change in administration, to build on what we had done.

At the very beginning of the project, we devoted one strand of our research to the study of alignment within the district. We chose to do this because 1) of the importance of alignment to the theory of systemic reform, 2) the district's strong emphasis on standards-based education, and, 3) interest on the part of MPS curriculum staff in having an external verification of the alignment of the newly adopted standards and assessments. We also felt that the study of alignment would give us a greater understanding of curriculum emphasis and assessments within the district.

By the 1998-1999 school year, Milwaukee Public Schools had important initiatives in place for an aligned system that were capable of concentrating effort on improved student achievement. The district had written content standards and grade-level expectations; set proficiencies and student requirements for grade 8 to grade 9 promotion (in 1999-2000) and for graduation from high school (in 2003-2004); and had an established standards-based assessment system, with related intervention programs and professional development. The district also was using a school-based decision-making model that allowed principals and their staffs some autonomy. The district was aligned to the degree that all of these components were working toward the same ends.

Over the previous six years, the district had been on a trajectory leading towards a standards-based system and increased student achievement. Such large-scale reform takes time to reach coherence among all district components. There was evidence in the four content areas—language arts, mathematics, science, and social studies—that progress had been made in aligning the important components of standards, curriculum, assessment, and professional development. Important steps towards a standards-based system included:

- Adoption of K-12 Teaching and Learning Goals (about 1994);
- Development of grade-level expectations in mathematics and science (from 1994);

- Mandated district graduation requirement and middle school proficiencies (1996);
- Implementation of a writing performance assessment;
- Milwaukee Urban Systemic Initiative (initiated 1996-97);
- MPS Standards and Grade-Level Expectations (1998);
- District-wide adoption of mathematics curriculum, grades 6-8 (1998-99); and,
- Professional development focused on new mathematics curriculum.

In November, 1998, the Milwaukee Board of School Directors approved the *Milwaukee Public Schools K-12 Academic Standards and Grade-Level Expectations* for language arts, mathematics, science, and social studies. Responding to demands of the state and building on work that had already been done, the MPS Division of Curriculum and Instruction developed these content standards. The standards and grade-level expectations in each content area were developed under the leadership of the content area curriculum specialist and with the help of teacher committees. The work in each content area took a different approach, based in part on what was already in existence. As a consequence, the formats for standards and grade-level expectations among the four content areas were different.

Initially, the features of the MPS standards-based system reviewed in our alignment study included the district's standards and grade level expectations, the State of Wisconsin Model Academic Standards, and the Wisconsin Knowledge and Concepts Examinations (WKCE) for grades 4, 8, and 10, and the grade 3 Wisconsin Reading Comprehension Test (WRCT). At the end of the 1998-1999 academic year, the alignment study was extended to gather data in three schools on how principals, teachers, and staff attended to the district's newly adopted standards and the state assessments in developing school plans and preparing for instruction.

Concurrently with the alignment analysis, we provided technical assistance to district staff. Our objective for the technical assistance was to inform district staff of our research findings, but also to gain an understanding from district staff of the pressures they were experiencing and how our research could be of use to them in their work. We reviewed the proposed mathematics performance assessment to be administered in March, 1999, and sent our comments on it to the performance assessment specialist for the Office of Research and Assessment on February 19, 1999. We also sent comments on the proposed science performance assessments on March 29, 1999. On February 24, 1999, the alignment research team presented to the director of the Division of Curriculum and Instruction Division and to curriculum specialists an outline of our thinking and of our study of alignment in the district, seeking their cooperation.

MPS Performance Assessments and Proficiencies

The services we provided to district staff gave us important insights into the district and its curriculum and assessment system. Milwaukee Public Schools had bought into the need for multiple measures of student performance and greatly valued having students complete performance assessments in addition to standardized norm-referenced tests. For example, in communications, middle school students must demonstrate

proficiency in three areas: writing, reading, and oral communication. In writing, all eighth grade students must produce four different samples of their writing (imaginative, expository, persuasive, and narrative). Teachers select these samples from students' work in grades 6, 7, or 8. Students are to demonstrate skills in reading on a formal reading assessment chosen and administered by the school, or on the state assessment. Teachers also maintain a district reading assessment instruction card on which each student's progress in reading is recorded. To assess oral communication skills, students are to present a 3-5 minute videotaped demonstration speech, persuasive speech, or an interview. Student proficiencies on each of these seven activities are rated on the basis of four proficiency levels—minimal (1), basic (2), proficient (3), and advanced (4). A total of 18 points are required for a student to be judged as having met the proficiency in communication. Under district guidelines, students must be given at least three opportunities to meet each proficiency, whether administered by classroom teacher, school, or district.

In mathematics, students must demonstrate their understanding of a range of algebra topics by including in their portfolio five examples of their work. Teachers are to judge students' knowledge of essential algebra topics using a four-point rubric. The essential topics—patterns of change, linearity, mathematical models and exponential functions, quadratic functions, and symbolic mathematics—are all included in the algebraic strand of the newly adopted middle grades mathematics program. Also, students will need to include in their mathematics portfolio one of the alternatives that demonstrate their proficiency in passing an on-demand mathematics assessment. They either need to satisfactorily pass the middle grades MPS mathematics performance assessment, the grade 8 Wisconsin Knowledge and Concepts Examination on mathematics, or grade 7 TerraNova mathematics test. As a final requirement for demonstrating their proficiency in mathematics, students are by the end of grade 8 to satisfactorily create a three-dimensional scale model, or package design, that demonstrates understanding of measurement, proportional reasoning, and geometric relations. As in communication, teachers are to judge students' level of proficiency using a four-point rubric and students are required to attain 21 points out of a possible 32.

This brief summary of our review of the MPS mathematics and science performance assessments illustrates one way we developed a deeper understanding of the activities within the district. Initially, we did a very deep analysis that included extensive feedback on the content was covered by each item and on the depth-of-knowledge required for a student to successfully complete the assessment activity. In judging the depth-of-knowledge, we used the same four-point scale used in the alignment analysis—recall (1), procedural/conceptual knowledge (2), strategic thinking (3), and extended thinking (4). The summary of our analysis (Figure 1) indicated that two of the items required students to apply reasoning and problem solving (Cellular Phone and Paper Problem), but that the other six items were judged to require a lower depth-of-knowledge than generally would be expected by a performance assessment and could be more easily assessed using multiple-choice items.

Figure 1. Analysis of MPS Mathematics Proficiency Performance Assessment for 1999.

Performance Assessment Title	Numeracy		Measurements		Graphs		Formulas		Equations		Ratios and Proportions		Statistics and Probability		Geometry		Data Analysis		Depth-of-Knowledge
	C	D	C	D	C	D	C	D	C	D	C	D	C	D	C	D	C	D	
Homeward bound	X	1	X	2							X	2							2
The Sales Trip	X	1	X	2							X	2							2
Barrel of Solvent	X	1	X	1			X	1											1
Cellular Phone	X	2			?	2			?	(3)	X	2							3
Paper Problem	X	3	X	3			X	1							?	(1)			3
ABCD Garden	X	3	X	2			X	1	?	2	?	(2)			?	(2)			2
School Dance	X	2							X	2									2
Recycling	X	1			X	1													1

C = Covered; D = Depth to which covered

? in a column C => A student may use knowledge from the content area or may not.

(3) in a column D => The depth-of-knowledge required if a student used knowledge from the designated content area.

Depth-of-Knowledge Levels

1 – Recall, memorization

2 – Conceptual/procedural knowledge

3 – Strategic thinking, reasoning, analysis

4 – Extended thinking, application to a real problem

Our analyses of the performance assessments gave us some understanding of the district's performance assessments and their quality. But more importantly, our analysis and subsequent discussions with the performance assessment specialist gave us a greater understanding of the role of performance assessment within the district, the role of the newly adopted standards, and of the district's operations. Incorporating performance assessment into the district's assessment system was important in providing teachers incentives to include similar experiences in their teaching. That is, performance assessment activities were considered good instructional activities. However, performance assessment instruments had to be structured very carefully so that there would be no surprises for the teachers as to what students were required to do on the assessment.

The impact of the time frame that the district worked under became very apparent as a result of our interaction with the performance assessment specialist and her feedback on our analysis of the performance assessment activities. She was under tremendous pressure to generate new performance assessments to meet the demands of the assessment system. Groups of teachers help write the activities, but sometimes there was not an opportunity to field test the activities, and, if any field-testing was done, it could only be done once. High-stakes performance assessments and assessments used for the middle school proficiencies and the high school graduation requirements had to be administered at least twice annually, once in the fall and once in the spring. A new assessment had to be developed for each new administration. We understood that our review was primarily helpful in identifying needed superficial changes rather than substantive changes that might require replacing an existing activity with a new activity. Scoring the performance assessment also became a financial and resource burden on the district. Volunteer teachers had to be paid for scoring sessions held on the weekends. In some cases, not enough teachers from a content area volunteered to score the assessments, so teachers from lower grades or from different content areas were used to do scoring. In 1999-2000, the district shifted the burden to the schools by having the principal at each school decide how best to score the performance assessments.

With hardly enough time to develop performance assessment, there was no time to do statistical analyses and consider psychometric qualities of reliability and validity. There also was no time to equate forms of the same performance assessment administered at different times. Under the circumstances, the performance assessment program was accomplishing a considerable amount with very limited resources and personnel. Much of what we learned about the assessment program could be gained through extensive interviews. What would prove more difficult for us as researchers to understand was the significant amount of effort expended to support the performance assessment and how difficult it was for the district to meet the demanding time scale imposed by the assessment system. We did make some suggestions for improving the performance assessment development procedure by providing enough activities for three or four forms of an instrument at one time and field-testing one or two items with each administration of an instrument. The performance assessment specialist read these suggestions, but did not see how they could be implemented in the near future.

In the summer of 1999, the performance assessment specialist worked with groups of mathematics and science teachers to write performance assessment activities. At the beginning of the institute, two researchers from our research team gave a presentation on alignment and went through an alignment process with the teachers by having them compare the state assessments and the performance assessments with the new standards. One goal of the training was for teachers to understand better how to think about the depth-of-knowledge required by an assessment activity and anticipated by a standard. In the training, groups of teachers coded the content standards and objectives measured by the assessment activity and compared the depth-of-knowledge level. The work of the teachers was incorporated into our alignment study and was compared to judgments made by the researchers.

An unanticipated finding resulted from this effort to train a group of MPS teachers to do an alignment study. There was about 67% correspondence between how researchers assigned a depth-of-knowledge code to standards and objectives and how teachers coded the standards and objectives. Rather than accept the difference only as a source of error, we analyzed the two sets of coding to determine whether there were any systematic differences. One noticeable difference was that teachers coded the depth-of-knowledge levels of some of the standards and objectives higher than the researcher. Based on our observations during coding and teachers' comments, we learned that teachers were coding some of the expectations (standards and objectives) *as they would teach the content topic to the student* rather than *as the knowledge they expected students to have*. For example, one grade 12 objective states:

> B.12.22
> Describe in words the relationship between the dependent and the independent variable in exponential growth or decay functions.

The researcher rated this objective with a depth-of-knowledge level of 2 (conceptual and procedural knowledge). However, the teachers rated this objective with a depth-of-knowledge level of 3 (strategic thinking). The researcher made a case that the objective expects students to know the concept of exponential growth or decay functions and that a simple statement that the independent variable is an exponent and the dependent variable is equal to an exponential function ($y = a^x$ or $y = a^{1/x}$) is required. Teachers rated the depth-of-knowledge as strategic thinking, indicating students would need to engage in significant reasoning to meet the objective, apparently because they thought of instructional activities that they would have students do in order for them to learn about exponential growth and decay functions. Such instruction would probably incorporate different forms of representations and have students solve problems whose solutions are exponential functions. This insight in the difference between how we as researchers thought about standards *as an outcome* and how teachers thought about standards *as instruction* was helpful to us as we interpreted the other alignment studies.

MPS Assessment System

Our effort to help the district think through the process of streamlining its assessment system reinforces the illustration of our approach to embedded research. During February and March 2000, a research team met to think through possible alternatives for the district's assessment system, taking into consideration the district's needs and goals. This research team raised issues and made suggestions to the district staff members we met with during this time. The goal was to develop at least two alternatives that could be presented to the deputy superintendents on March 30 and then to groups of principals during the first week in April. Our role was to provide advice and expertise on assessments and accountability. The final decision, of course, was the district's.

Through our embedded research approach to specified modifications of the district's assessment system, we increased our understanding of the change in emphases in the district and of its current priorities. For example, in our first meeting with district staff, February 1, 2000, staff members identified goals for the assessment system. Our role was to help clarify goals raised rather than to recommend goals. From this experience, we learned about changes in emphases in assessment, at least in the district administration. One goal was for the district assessment system to be aligned with *both* the MPS and Wisconsin standards rather than with just the MPS standards. The ensuing discussion indicated that up until now assessments, such as performance assessments, were being used to drive instruction. Now there was more agreement that assessments should be selected and developed to match instruction.

From February through March, we engaged in a process to help the district think through alternatives for a district-wide assessment system responsive to the new legislative mandate of a high school graduation test and the district's capacity. The timeline was accelerated because of the need to present a recommendation to the Board of School Directors by May so that changes to the assessment system could be implemented in the 2000-2001 academic year—the first group of grade 9 students who will be required to take the newly mandated high school graduation test in 2003-2004. For us as researchers, this timeline was too short to consider all of the possibilities. We would have liked to complete our study on the proficiencies in order to inform the process. We also felt that it would make better sense to phase in some of the modifications of the assessment system. This again demonstrated the differences between our research time frame and the district time frame. What we did do was to develop a process that defined alternatives for the assessment system by the end of March that could be presented to groups of principals in April and then put in a form that could be presented to the Board of School Directors in May. During this time, staff from the Office of Research and Assessment conducted focus groups of school staff, including teachers, curriculum coordinators, and principals _____.

The existing assessment system included multiple measures and a variety of assessments across content areas and grade levels. In the 1999-2000 school year, the district assessment consisted of state-mandated tests, proficiency assessments,

performance assessments, and portfolios. The middle grades students were assessed on the proficiencies in communications, mathematics, science, and research. These have been described above. Students in grades 4, 8, and 10 were required by the state to take the Wisconsin Student Assessment System (WSAS) Knowledge and Concept Examinations. Grade 3 students were required to take the Wisconsin Reading Comprehension Test (WRCT). In addition to these assessments, an MPS mathematics proficiency assessment and a writing proficiency assessment were administered to students in grades 11 and 12 as a high school graduation requirement. MPS performance assessments were given in writing, science, fine arts, and oral communications. In the spring, students in grades 4 and 5 were required to write an essay to a specific prompt. Science performance assessments were administered to students in grade 5; and grades 9, 10, and 12. Each high school had a plan that assessed about one-third of the students in these grade ranges each year. Each school was required to administer either a fine arts assessment or an oral communication assessment. High schools and middle schools determined when the fine arts or oral assessments were administered. Elementary schools were to administer these assessments to students in grade 4 or 5. Students who did not pass the proficiency assessments in these two content areas could complete portfolio assessments in writing and mathematics as an alternative means for meeting the district graduation requirement.

During February and March, we had a series of meetings with staff members from the Office of Research and Assessment, Audit Services, and Special Services (special education). The last meeting in March included the two deputy superintendents, the director of Educational Services, and the director of the Office of Curriculum and Instruction Division, at which alternative assessment options were identified and discussed. The alternative that received the most acceptances at the time was labeled a value-added assessment system. This would include administering standardized norm-referenced examinations in each grade from grade 1 to grade 10 in reading, writing, language arts, mathematics, science, and social studies. The district would identify a test to be administered in the years that the state assessments were not administered. In addition, performance assessments would be administered as part of the high school graduation requirements in grades 11 and 12 in writing and mathematics, along with the high school graduation test in language arts, mathematics, science, and social studies. These assessments would constitute external measures, assessments developed and scored at the district or state level. The external assessments would be accompanied by internal assessments, those assessments administered and scored by teachers. The internal assessments would vary by content area. For reading, for each grade K-8, teachers would verify student's reading level using one of a number of available standard instruments. For the other content areas, teachers would be required to check on students' progress in learning to use classroom assessments based on standards. Exactly what would constitute classroom assessment based on standards has not been specified. These could be similar to the algebra requirements for the current middle grades proficiencies—that is, a curriculum event, project, or activity critical for students in meeting the standards that would require teachers' verification of students' satisfactory completion. A set of curriculum-based assessments would be specified in writing for grades K-8 and in language arts, mathematics, science, and social studies for grades K-12.

The new Wisconsin legislation requires school districts to specify the criteria for granting a high school diploma that includes high school graduation test scores, pupil academic performance, and recommendations of teachers. Within these guidelines, the legislation defers to the district to identify precisely how each of these criteria is to be met. The state has specified similar criteria for promotion to grade 5 and grade 9. What the criteria for MPS should be was an issue incorporated into the discussion of the assessment system. At the time this report was written, a firm decision had not been reached on what criteria MPS should specify. The district was considering requiring that students receive a satisfactory score in each content area on at least one of three criteria—the state test (the meaning of satisfactory would have to be defined), demonstrate academic performance as determined by the classroom assessment based on the standards, and growth in achievement as judged by teachers.

Another insight we gained was that the value-added concept was now more acceptable within the district and was reported by the district staff as being supported by the president of the Board of School Directors. We probed further about how those in the district were thinking about the value-added approach and how they would distinguish between an assessment system that provided information that would improve instruction and, consequently, student learning and an assessment system that tracked students' annual gains. The district staff felt that the sentiment was towards a value-added system that would improve instruction as well as track students' yearly progress. Two months later, as the process converged on primarily one alternative, how value-added principles should be used in an assessment system became a point of contention between district staff and our research team. We made the case that value-added procedures were most appropriate for school accountability, but not for student accountability from grade to grade. We supported the administration of standardized norm-referenced tests in each grade in six content areas in order to make more accurate judgments on the improvements of schools. The district staff was very interested in tracking individual student growth. The reason they wanted annual testing with standardized norm-referenced tests was to track gains by individual students from year to year. At the time of this report, we had provided the district staff with reasons for why the use of individual gain scores was less reliable—large standard error of measurement in individual gain scores, easily corruptible, and differential advantages for some students over other students. The district staff felt very strongly that student progress in learning should be included as one of the three criteria to be considered for high school graduation and grades 5 and 9 promotion. The research team is discussing further the ways in which student growth could reasonably and reliability be used as a defendable criterion. (We were to learn at an April 13 meeting that the name for the external assessment system was changed from "value-added" to "longitudinal" assessment.)

A third insight we gained during the February 1 meeting was a possible shift in the district's thinking about offering multiple opportunities for students to demonstrate what they had learned. The district's history in the past five years of using a variety of measures, including performance assessments, standardized norm-referenced tests, and portfolio activities incorporated in the middle grades proficiencies, was described by one

district staff member as more a reflection of the culture and shared beliefs in the district than a statement of policy. But it was this person's perception that the new administration did not place as much weight on using multiple measures. As the process evolved, most alternatives for the assessment system incorporated both external and internal measures of student learning. The discussion supported our sense that district staff valued more than one measure. However, the recommendation for the graduation criteria advanced by the district staff at the March 30 meeting indicated that the concept of multiple measures was not considered as important or relevant. The district staff recommended that "[a] student will have the opportunity to demonstrate proficiency in three different ways: 1) Test Results, 2) Academic Performance, or 3) Recommendations of Teachers." The test results are to be based on student performance on the state's high school graduation test. Academic performance is academic growth in each subject area as measured by the "MPS value-added assessment." Recommendations of the teachers are the grade point averages (GPAs) in each subject area to be determined in part by classroom assessments. Members of our research team argued at the March 30 meeting and subsequently that the first two criteria were not distinct measures but highly correlated—scores on the high school graduation test and academic growth determined by using standardized norm-referenced test results. Instead, our research team recommended that the three criteria be:

1. An acceptable score on the high school graduation test;
2. Satisfactory performance on classroom assessments, based on the standards, administered in grades 9, 10, 11, and 12; and,
3. A decision by a panel of three school staff members as to whether a student had made adequate progress in the later high school years, taking into consideration improved performance on standardized tests, attendance, and other factors.

One concern raised by the district staff was that the teacher's recommendation criterion be based on objective measures that could be validated. We felt that using a panel would address this issue, but organizing a panel of three persons to evaluate each student at the end of grade 12 was viewed as problematic by the district staff.

How and whether these points will be resolved is still in question at this point in time. There seems to be agreement that the district needs to have annual testing to strengthen the accountability system. This would address one of our concerns that there have been insufficient data to judge whether the district and schools are improving as judged by student learning. We have tried to increase district staff's understanding of concepts such as value-added and multiple measures in assessments. We also have provided assistance to some professionals in the district on how to conduct alignment analyses and tools for judging the relation of assessment activities to standards and grade-level expectations. To the degree that we have been successful in working with district staff in these ways, we can see that the analytic capacity of district staff is improving.

Summary

This paper has attempted to illustrate one line of inquiry, what we are calling embedded research, as we work with the Milwaukee Public Schools and study systemic change. This line of research has evolved from alignment analyses to assessment system design. Through the process of helping the district address important issues, we have endeavored to develop district analytic capacity. Over the course of the two years, we have worked with MPS and through one change in administration. Our work has primarily been with district staff members in the Office of Research and Assessment, the Division of Curriculum and Instruction, and Technical Services. We have gathered data and interacted with personnel in a few schools and are currently working more intently with staff in two middle schools. There are indications that our work with the district has been valued, at least by some, by their willingness to work with us and engage in thinking through some of the most pressing issues for the district, such as the modifications to the assessment system. There is evidence of trust building and collaboration between some district staff members and our research team.

Clune's model of embedded research as systemic capacity building has five components: a theoretical base, inputs, building understanding, outputs, and practical feasibility. The embedded research illustrated in this example was guided by a theory of systemic reform that requires the alignment of system components. The research also incorporated system and school accountability and assessment principles, including consequential validity and the use of multiple measures. The inputs from the district included newly adopted standards and grade-level expectations; an assessment system that incorporated a variety of measures and that drove instruction (but did not have the resources to develop psychometrically sound instruments); a strong Middle School Principals Collaborative and coherence in the middle grades; a newly adopted district-wide mathematics curriculum for the middle grades; a data warehouse system under development; and a new administration that is decentralizing the district by moving nearly all decision responsibility to the schools.

Our research has improved our understanding of the district in subtle ways. The extended and continuing interaction we have had with district staff in rethinking the assessment system has helped us understand better how key ideas are being interpreted. For example, the value-added concept was used by at least some district staff both to refer to school accountability and student accountability. This became more evident when it was seen that the same assessment instruments and analyses could be used for both. For value-added measures to be successfully incorporated into an accountability system, district decision makers will need a better understanding of this approach to data analysis. District staff members continue to value alternative measures of student achievement and how performance assessment has influenced instruction, particularly in writing and mathematics. This is something that staff members wanted to retain in the revised assessment system, but for these measures to be internal rather than external. In addition to requiring more resources than are available, what is missing from the assessment system is some objective measure of student progress.

Practical constraints have also informed our developing understandings. As noted above, the district was unable to devote sufficient resources to develop psychometrically

sound performance assessments with different equivalent forms. Used as one measure of a collection of criteria for judging proficiency, there was general acceptance in the district of the performance assessments and a perception, at least by some, that these had a strong influence on what teachers did with their students in classrooms. Performance assessments had been supported by teachers for a number of years. Although the new administration supports high standards for all students, the administration is working to transfer 90% of the district's budget to the control of the school principals. This, along with other budget constraints, will further limit the amount of assessment development the district can support.

It is too soon to determine the outcomes for the district from our embedded research approach. The form the new assessment system assumes and the criteria used for judging graduation and promotion requirements will be important indications of the impact we have had within the district. We do believe district staff members are thinking more deeply about value-added and multiple measures, but we are not sure whether other pressures, such as cost and manageability, will be of greater concern. What is more evident now than it was a year ago is that our intermediate goal of being critically positioned and engaged in helping the district think through some of its most pressing issues is being realized.

One crucial consideration in any research is generalizability. Embedded research, as we are applying it in our work with MPS, gives us a deeper understanding of the district's operations than we would have if we were external observers or data gatherers. Understanding fully how district staff members are using terms, such as value-added and multiple measures as well as standards and alignment, has required a number of interactions with district staff over several occasions. Through the act of joint problem solving, as researchers we not only hear what district staff members say, but we understand more fully the thinking underlying their words. As researchers, we continually have to step back and reflect on what we are doing, what the theoretical bases for our work are, and what the logic-of-action of the district is. Utilizing a multidisciplinary research team, where not all team members are engaged in all research projects, is critical. Drawing on these multiple perspectives facilitates an objective review of one research group's efforts by others, which serves to enhance the learning and improve the technical assistance and responsiveness of our research team. At a minimum, the findings from our embedded research work will serve as a detailed case study of systemic change. The potential to generalize from these findings will rest on how successful we are at identifying reasons for system change and how components common to any large urban district interact to further or retard this change.

References

Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *Journal of the Learning Sciences, 2* (2), 141-178.

Century, J. R. (1999). Evaluators' roles: Walking the line between judge and consultant. In N. L. Webb (Ed.), *Evaluation of systemic reform in mathematics and science. Synthesis and proceedings of the Fourth Annual NISE Forum.* (Workshop Report No. 8.) Madison: University of Wisconsin, National Institute for Science Education.

Chen, H-T. (1990). *Theory-driven evaluations.* Newbury Park, CA: Sage Publications.

Clune, W. H. (2000). *Embedded research on systemic reform and the design experiment: Similarities and dissimilarities.* Paper presented at the American Educational Research Association Annual Meeting, New Orleans, April 24-28.

Clune, W. H. (1998). *Toward a theory of systemic reform: The case of nine NSF Statewide Systemic Initiatives.* (Research Monograph No. 16). Madison: University of Wisconsin, National Institute for Science Education.

Clune, W. H., & Webb, N. L. (1997). *Center for the study of systemic reform in Milwaukee Public Schools,* A proposal to the Joyce Foundation. Madison: University of Wisconsin, Wisconsin Center for Education Research.

Consortium for Policy Research in Education. (1991). *Putting the pieces together: Systemic school reform* (CPRE Policy Briefs). New Brunswick, NJ: Rutgers, The State University of New Jersey, Eagleton Institute of Politics.

Heck, D., & Webb, N. L. (1996). *Purposes and issues of systemic evaluation in education as reflected in current evaluations and literature.* Unpublished manuscript.

Office of Research & Assessment. (1998) *Milwaukee Public Schools: Programs and initiatives.* Milwaukee, WI: Milwaukee Public Schools.

Newmann, F. (1993). Beyond common sense in educational restructuring: The issues of content and linkage. *Educational Researcher, 22* (2), 4-13, 22.

Smith, M. S., & O'Day, J. (1991). Systemic school reform. In S. H. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing* (pp. 233-267). *Politics of Education Association Yearbook.* (1990). London: Taylor & Francis.

U.S. Congress, House of Representatives. (1994). *Improving American's Schools Act.* Conference Report to accompany H. R. 6 Report 103-761. Washington, DC: U. S. Government Printing Office.

Webb, N. L. (forthcoming). Alignment. In N. L. Webb, J. R. Century, N. Dávila, D. Heck, & E. Osthoff (Eds.), *Evaluation of systemic reform in mathematics and science.*

Zucker, A. A., Shields, P. M., Adelman, N. E., Corcoran, T. B., & Goertz, M. E. (1998). *A report on the evaluation of the National Science Foundation's Statewide Systemic Initiatives (SSI) Program.* Menlo Park, CA: SRI International.